The Validity of Lexicon-based Emotion Analysis in Interdisciplinary Research

Emily Öhman
Waseda University
ohman@waseda.jp

Abstract

Lexicon-based sentiment and emotion analysis methods are widely used particularly in applied Natural Language Processing (NLP) projects in fields such as computational social science and digital humanities. These lexicon-based methods have often been criticized for their lack of validation and accuracy – sometimes fairly. However, in this paper, we argue that lexicon-based methods work well particularly when moving up in granularity and show how useful lexicon-based methods can be for projects where neither qualitative analysis nor a machine learning-based approach is possible. Indeed, we argue that the measure of a lexicon's accuracy should be grounded in its usefulness.

1 Introduction

Lexicon-based sentiment analysis is probably the simplest approach to determining the polarity or emotional content of a text. At the core, it is simply comparing the lemmatized tokens in a text being analyzed to the lemmas in the lexicon and assigning them a score accordingly. Some models go a step further and try to include valence-shifters in the final polarity measure, but most commonly lexicon-based sentiment analysis relies on bag-of-words approaches (see e.g. Taboada et al. (2011)).

Lexicon-based methods are commonly contrasted with machine learning-based methods (Nguyen et al., 2018; Kaushik and Mishra, 2014; van Atteveldt et al., 2021). Machine learning is typically context sensitive and can be combined with large language models for a fairly accurate picture of the linguistic content of a text. Machine learning models typically perform much better than lexicon-based models on sentiment analysis tasks when comparing traditional evaluation metrics (Kaushik and Mishra, 2014; González-Bailón and Paltoglou, 2015; Dhaoui et al., 2017; van Atteveldt et al., 2021).

However, there are two issues with comparing machine learning, or data-driven, and lexicon-based models. The first is that an accurate machine learning model needs labeled data for training, validation, and testing. Labeling or annotating data generally requires at least three human annotators who need to be compensated for their work. Therefore machine learning datasets can quickly exceed the budget of many projects or be flat out impossible to conduct properly especially for early career researchers with smaller amounts of grant money to spend on such tasks (see e.g. Gatti et al. (2015)).

The second issue, which is rarely discussed, is that these evaluation metrics are not really comparable due to how emotion and sentiment scores are assigned and calculated using these different methods. Naturally, any approach needs to be evaluated, but in practice it is much harder to accurately evaluate the output of a lexicon-based model. Instead, the focus should be on usefulness of the output accompanied by a sanity check of the results. The validation issue is discussed in detail in section 3.

It is rare to see lexicon-based methods used for sentiment analysis in NLP papers. Conversely, it is fairly common to see them used in interdisciplinary projects. In some fields, there are very few scholars willing to review interdisciplinary papers, and even fewer who have the expertise to properly make judgments on the methodology. This can lead to some papers being accepted that have dubious methodology or other being rejected because they are too technical. This is something that affects interdisciplinary fields the most.

In the following pages we present an overview of interdisciplinary sentiment analysis practices and common criticism against different methods. We also discuss the issue with the evaluation of lexicon-based sentiment analysis projects and offer

Proceedings of the Workshop on Natural Language Processing for Digital Humanities (NLP4DH), pages 7–12
co-located with ICON-2021, December 16-19, 2021, Silchar, India. ©2021 NLP Association of India (NLPAI)
https://doi.org/10.26615/978-952-94-5833-2_002

some preliminary solutions while making a case for lexicon-based sentiment analysis in interdisciplinary projects.

2 Background

2.1 The Creation and Validation of Emotion Lexicons

There are a few different ways of creating emotion lexicons, however, typically some type of emotion dictionary is used to extract relevant lexical items (Mohammad and Turney, 2010). There are many ways of annotating for emotions. Annotators might be asked to annotate for emotion evocation or emotion association, which can result in very different results. Mohammad and Turney (2013) found that annotating for emotion association resulted in more reliable annotations. Annotator fatigue is also very common, especially when annotating for emotions or sentiments (Mohammad, 2016; Öhman, 2020a) and therefore the method of annotation also has a significant impact on the quality of annotations (Kiritchenko and Mohammad, 2017).

Nonetheless, these lexicons are carefully constructed and inter-annotator agreement scores are carefully evaluated. Noisy annotations and even noisy annotators are often removed before compiling the final lexicon. The reality is simply that human annotators do not always agree on an annotation, and when the annotation task is emotion annotation, disagreements are even more common (Strapparava and Mihalcea, 2007; Andreevskaia and Bergler, 2007; Wiebe and Riloff, 2005). A typical inter-annotator agreement percentage is around 70% but can be much lower than that (Bermingham and Smeaton, 2009; Ng et al., 1999).

2.2 Common complaints

Particularly in computational social sciences and digital humanities the use of black-box or black-box-like tools is fairly common (Lazer et al., 2020; Gefen et al., 2021). Often this tool is LIWC (Linguistic Inquiry and Word Count) (Pennebaker et al., 2001) (especially in Computational Social Sciences) (Puschmann and Powell, 2018). In essence LIWC is an emotion lexicon and part-of-speech tagger (with other additional features).

LIWC itself has attracted criticism beyond the typical complaints against lexicon-based methods (Puschmann and Powell, 2018), not because the lexicon is any worse than any other emotion lexicon, but because the creators of LIWC not only claim that LIWC can detect emotions in text, but that it can also accurately identify a person's psychological state (Tausczik and Pennebaker, 2010). Indeed, this approach has been attempted in digital humanities research too (Boyd, 2017).

Another famous example is the Syuzhet package for R (Jockers, 2015). In essence, Syuzhet accepts a text (a novel) as input, compares the words in the text to those in the lexicons available to it, and outputs different visualizations of sentiment polarities in the text along the narrative path. Syuzhet has received multiple complaints ranging from statistical issues to the very practical issue of valence shifters not being taken into consideration and an over-reliance of word-occurrence (Swafford, 2015, 2016).

In their excellent work Data-sitter's club, DH project of the year 2019, Bowers and Dombrowski (2019) exemplify the problem with sentiment analysis in digital humanities by comparing some common programs for sentiment analysis such as Vader-Sentiment (Hutto and Gilbert, 2014) and TextBlob (Loria, 2018). They look at individual sentences and compare their human judgment with the judgments of these programs that are lexicon-based. Their conclusion is that sentiment analysis, particularly lexicon-based, is highly inaccurate.

These examples highlight the issue with lexicon-based approaches for emotion detection. Even the most accurate emotion lexicon still just counts words in the target text. To make claims beyond the occurrence of emotion-associated words is misguided at best and disingenuous at worst. However, this does not mean that there is something wrong with the lexicons themselves. It also does not mean that these lexicons can not be useful for sentiment and emotion analysis.

2.3 Comparison between lexicon-based and data-driven approaches

González-Bailón and Paltoglou (2015) compared the performance of several off-the-shelf sentiment lexicons to a machine learning approach. They concluded that lexicon-based methods performed comparable to machine learning, but that the accuracy of lexicon-based methods suffered more when the content was more diverse or informal. They calculated the accuracy by comparing to human annotators. Their final recommendation is to use machine learning.

van Atteveldt et al. (2021) do something similar, but they also add a layer of complication to their validity measures in that they translate their content to English from Dutch in order to use many off-the-shelf lexicons. Nonetheless, their results too, when compared to human coders, suggest that machine learning approaches have a much higher accuracy than lexicon-based ones.

3 The Issue with Validation

Validation is a complex matter, especially when we are evaluating lexicon-based emotion analysis projects in digital humanities projects that analyze more than the performance of a model. The approaches to validation differ greatly between fields. Some simply trust that LIWC does what it says it does (as evidenced by over 13,000 combined citations for the tool on Google Scholar[1]) and complete no further validation or sanity check, some look at a few individual data points and annotate or evaluate this themselves by comparing directly with the results from their model's output.

Naturally, for a lexicon-based method to be considered useful, the output should be close to a qualitative evaluation by a human. However, as already discussed, humans rarely agree on the emotional content of a word, sentence, or paragraph, so whether we are using annotated data for lexicons or to train machine learning models, it is hardly surprising that the output mirrors the confusion in the data. Only with highly disjoint categories and quality annotations devoid of noise is it possible to get high accuracies with classification tasks (see e.g. Demszky et al. (2020) and Abdul-Mageed and Ungar (2017)).

van Atteveldt et al. (2021) suggests that to measure the validity of a lexicon-based approach, one should manually annotate at least 100 data points, but ideally 300 for accurate Krippendorf's α scores. But if we are examining literary works or political party manifestos, this is not really possible as the unit of evaluation is typically a full document and we rely on composite scores for a unit at a coarser granularity than what the model is evaluating at. Furthermore, if a lexicon offers a range of scores beyond 0 and 1, such as intensity scores between 0 and 1 for each lexical item it is quite difficult to source these human annotations as humans are notoriously bad at rating scales and the emotion intensities in the NRC Emotion Intensity lexicon (Mohammad et al., 2018), for example, were obtained using best-worst-scaling (Kiritchenko and Mohammad, 2017), something which is typically not feasible to conduct for small batches of test data.

The problem with the suggested validation steps of such results is that (1) they work best for evaluating binary or ternary **sentiment** categorization, and (2) they evaluate computational approaches against human annotations. In some cases the latter makes sense. If we are analyzing tweets or other short messages for sentiment or emotion it makes sense to look at the assigned emotion scores at sentence- or message-level and these are relatively easy to compare against human annotations. However, the manual annotation for emotions typically produces different output than what lexicon-based models do and direct comparisons can be difficult in the best of circumstances. If we are working with emotion analysis with six, eight, or even more categories or emotion intensities instead of binary categories the results become even more complex and more difficult to compare against human annotations or machine learning approaches (Öhman, 2020b). The expectation still seems to be to follow the guidelines of binary sentiment analysis validation at sentence-level even when using multiple emotion categories for emotion intensity at document level.

If we are analyzing the emotional intensity of each named emotion in the content of speeches, party manifestos, or romance novels we are typically doing this analysis for chunks of 3,000-10,000 words. Following the validation guidelines of van Atteveldt et al. (2021), i.e. annotating a minimum of 100 units would mean manually annotating the emotional intensity of at least 300,000 words by 2-3 annotators. This amount of annotations is not even necessary to train a machine learning model. The next best thing then becomes annotating 2-3 chunks at sentence-level and calculating a composite score of the human annotations as well. This would typically result in at least 1,000 manual annotations which is more than enough to calculate Krippendorf's α accurately. However, this still leaves us with the issue of how human annotators would reliably be able to annotate for emotional intensity as there is little **intra**-annotator agreement, let alone inter-annotator agreement when annotating for scale. Furthermore, most lexicon-based models would likely score a sentence with two words expressing *sadness* as having twice the

[1] Accessed on November 12, 2021

sadness of a sentence that contained only one such word, but when a human annotator manually annotated that sentence, it is quite likely that they would only mark it as containing *sadness* in general, again making direct comparisons more difficult.

4 Proposed Solution & Use case

The first step would be to stop calling what lexicon-based methods do *emotion* or *sentiment analysis* and refer to it as analyzing the distribution of emotion-laden words. This is a much more accurate description of what lexicon-based methods actually do, especially when contrasted with what machine learning based methods do. If lexicon-based approaches are used together with statistical significance calculations, we can show that there are significant differences between the use of words associated with specific emotions in two comparable texts. This is in itself a demonstration of usefulness.

Such an approach also minimizes the need for adjusting the results for valence-shifters. If we are evaluating the use of emotion words in novels, whether an emotion word is negated or not is not as relevant because in this case authors choose their words to evoke specific emotions in the reader and thus such an approach is excellent for measuring tone and mood in text. It might even be argued that such an approach to tone and mood is going to result in more relevant results than a machine learning approach would as it might be easier to access the author's intent rather than the surface of the words. Word choice by literary authors has been shown to affect the mood of the novel significantly (McCormack, 2006; Ngai, 2005).

Another domain where words are carefully chosen is politics (Riggins, 1997; Orwell, 1946). Comparing the content of two political manifestos the distribution of emotion words when combined with statistical significance testing, can show us what type of emotion words are used more in each of the manifestos, and whether the difference is statistically significant. If the differences are statistically significant, the fact that the results indicate that one party used different words to evoke different emotions or words of different intensity is a useful finding. As a side note, especially when using off-the-shelf general purpose lexicons, it is a good idea to stick to formal single-domain texts in order to maximize the validity of the results as suggested by the results of González-Bailón and Paltoglou (2015).

The solution is to establish an evaluation metric for lexicon-based methods that focuses on usefulness rather than accuracy. A part of this usefulness measure would include doing some type of sanity check or validation comparing to human impressions of the text, but would take into account the different outputs of the lexicon-based model and the human annotations. A part of this validation can indeed use Krippendorf's α scores to check for inter-annotator agreement between the human annotators, as these annotators would have annotated the text in comparable ways. The comparison between the outputs of the model and the human annotators requires other metrics to determine usefulness or even traditional accuracy depending on what exactly the model outputs.

4.1 Use case

We have achieved the best results by letting the model add word scores that are then combined at document-level for a document emotion-word intensity score for each emotion using Plutchik's 8 core emotions sans *surprise* (Plutchik, 1980) as *surprise* is notoriously difficult to detect in text, particularly at sentence-level or finer granularity (Alm and Sproat, 2005). The human annotators (at least 2, but ideally 3) annotate a few select representative documents at roughly sentence level by simply marking the sentence as containing the emotions in the annotation scheme. Although humans annotate for the binary existence or non-existence of the particular emotions, the results are far more reliable than if they were to annotate for intensity (Kiritchenko and Mohammad, 2017). The results also correlate highly with those of the intensity-scores both in terms of absolute numbers and proportional distribution of emotions.

In one instance we examined Finnish political party manifestos (Koljonen et al., forthcoming). We used a straight-forward emotion intensity lexicon that had been adjusted for political data and the Finnish language (Öhman, forthcoming) to get composite scores for nearly 1000 party manifestos that were on average around 20,000 tokens in length. Using linear regression to analyze statistical significance showed that the main difference between different parties, manifesto types, and eras, was that although populist parties used the same amount of emotion words as other parties, the words they used were of significantly higher intensity.

We confirmed our findings by having three annotators annotate three manifestos, by different political parties and different eras, manually at approximately sentence-level by marking that sentence as expressing or not expressing any of the emotions in our scheme. We calculated inter-annotator agreement using Krippendorf's α which was on par with other emotion annotation tasks. We then compared that score to the compound score adjusted for word count from our model. The values per emotion were nearly identical. It was not possible to do a direct inter-rater agreement calculation, but comparing the distribution of emotions, the values were again nearly identical for all the target manifestos.

Comparing the manual annotations to the output of the lexicon would not have yielded any useful metrics. However, the significance calculations show that there was valuable undiscovered information in the data that we could access with emotion lexicons.

5 Concluding Discussion

In this opinion paper we have tried to justify the use of lexicon-based emotion analysis, particularly in interdisciplinary research. There is little doubt that data-driven methods such as machine learning are typically the best choice when aiming for accuracy, however, there are projects and approaches where lexicon-based methods fare equally well, and sometimes are even more suitable for the task than machine learning. We hope this paper initiates a discussion in particular about the process of validating results from lexicon-based approaches in a way that would recognize the usefulness of lexicon-based approaches for specific types of text commonly used in digital humanities.

References

Muhammad Abdul-Mageed and Lyle Ungar. 2017. Emonet: Fine-grained emotion detection with gated recurrent neural networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 718–728.

Cecilia Ovesdotter Alm and Richard Sproat. 2005. Emotional sequencing and development in fairy tales. In *International Conference on Affective Computing and Intelligent Interaction*, pages 668–674. Springer.

Alina Andreevskaia and Sabine Bergler. 2007. CLaC and CLaC-NB: Knowledge-based and corpus-based approaches to sentiment tagging. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, SemEval '07, pages 117–120, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wouter van Atteveldt, Mariken ACG van der Velden, and Mark Boukes. 2021. The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, 15(2):121–140.

Adam Bermingham and Alan F Smeaton. 2009. A study of inter-annotator agreement for opinion retrieval. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 784–785.

Katherine Bowers and Quinn Dombrowski. 2019. Katia and the sentiment snobs. In *The Data-Sitters Club*.

Ryan L Boyd. 2017. Psychological text analysis in the digital humanities. In *Data analytics in digital humanities*, pages 161–189. Springer.

Dorottya Demszky, Dana Movshovitz-Attias, Jeongwoo Ko, Alan Cowen, Gaurav Nemade, and Sujith Ravi. 2020. Goemotions: A dataset of fine-grained emotions. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4040–4054.

Chedia Dhaoui, Cynthia M Webster, and Lay Peng Tan. 2017. Social media sentiment analysis: lexicon versus machine learning. *Journal of Consumer Marketing*.

Lorenzo Gatti, Marco Guerini, and Marco Turchi. 2015. Sentiwords: Deriving a high precision and high coverage lexicon for sentiment analysis. *IEEE Transactions on Affective Computing*, 7(4):409–421.

Alexandre Gefen, Léa Saint-Raymond, and Tommaso Venturini. 2021. Ai for digital humanities and computational social sciences. In *Reflections on Artificial Intelligence for Humanity*, pages 191–202. Springer.

Sandra González-Bailón and Georgios Paltoglou. 2015. Signals of public opinion in online communication: A comparison of methods and data sources. *The ANNALS of the American Academy of Political and Social Science*, 659(1):95–107.

Clayton J. Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *ICWSM*. The AAAI Press.

Matthew L Jockers. 2015. Syuzhet: Extract sentiment and plot arcs from text. *blog post*.

Chetan Kaushik and Atul Mishra. 2014. A scalable, lexicon based technique for sentiment analysis. *arXiv preprint arXiv:1410.2265*.

Svetlana Kiritchenko and Saif Mohammad. 2017. Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 465–470. Association for Computational Linguistics.

Juha Koljonen, Emily Öhman, Mikko Mattila, and Pertti Ahonen. forthcoming. Strength and intensity of sentiments and emotions in party manifestos: Finland, 1945 to 2019.

David MJ Lazer, Alex Pentland, Duncan J Watts, Sinan Aral, Susan Athey, Noshir Contractor, Deen Freelon, Sandra Gonzalez-Bailon, Gary King, Helen Margetts, et al. 2020. Computational social science: Obstacles and opportunities. *Science*, 369(6507):1060–1062.

Steven Loria. 2018. textblob documentation. *Release 0.15*, 2.

Thomas McCormack. 2006. *The fiction editor, the novel, and the novelist*. Paul Dry Books.

Saif Mohammad. 2016. A practical guide to sentiment annotation: Challenges and solutions. In *WASSA@ NAACL-HLT*, pages 174–179.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. SemEval-2018 task 1: Affect in tweets. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 1–17. Association for Computational Linguistics.

Saif Mohammad and Peter Turney. 2010. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34.

Saif M. Mohammad and Peter D. Turney. 2013. Crowdsourcing a word-emotion association lexicon. *Computational Intelligence*, 29(3):436–465.

Hwee Tou Ng, Chung Yong Lim, and Shou King Foo. 1999. A case study on inter-annotator agreement for word sense disambiguation. In *SIGLEX99: Standardizing Lexical Resources*.

Sianne Ngai. 2005. *Ugly feelings*, volume 6. Harvard University Press Cambridge, MA.

Heidi Nguyen, Aravind Veluchamy, Mamadou Diop, and Rashed Iqbal. 2018. Comparative study of sentiment analysis with product reviews using machine learning and lexicon-based approaches. *SMU Data Science Review*, 1(4):7.

Emily Öhman. 2020a. Challenges in Annotation: Annotator Experiences from a Crowdsourced Emotion Annotation Task. In *Digital Humanities in the Nordic Countries 2020*. CEUR Workshop Proceedings.

Emily Öhman. 2020b. Emotion annotation: Rethinking emotion categorization. In *DHN Post-Proceedings*, pages 134–144.

Emily Öhman. forthcoming. SELF & FEIL: Sentiment and Emotion Lexicons for Finnish.

George Orwell. 1946. Politics and the english language.

James W Pennebaker, Martha E Francis, and Roger J Booth. 2001. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71(2001):2001.

Robert Plutchik. 1980. A general psychoevolutionary theory of emotion. *Theories of emotion*, 1:3–31.

Cornelius Puschmann and Alison Powell. 2018. Turning words into consumer preferences: How sentiment analysis is framed in research and the news media. *Social Media+ Society*, 4(3):2056305118797724.

Stephen Harold Ed Riggins. 1997. *The language and politics of exclusion: Others in discourse*. Sage Publications, Inc.

Carlo Strapparava and Rada Mihalcea. 2007. Semeval-2007 task 14: Affective text. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 70–74.

Annie Swafford. 2015. Problems with the Syuzhet package. *Anglophile in Academia: Annie Swafford's Blog*.

Joanna Swafford. 2016. Messy data and faulty tools. *JSTOR*.

Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. 2011. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307.

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54.

Janyce Wiebe and Ellen Riloff. 2005. Creating subjective and objective sentence classifiers from unannotated texts. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 486–497. Springer.